OXFORD (GIGA)^n SCIENCE

GigaScience, 2022, **11**, 1–15
DOI: 10.1093/gigascience/giac082
RESEARCH

Statistical quantification of confounding bias in machine learning models

Tamas Spisak *

Center for Translational Neuro- and Behavioral Sciences, Institute for Diagnostic and Interventional Radiology and Neuroradiology, Center University Hospital Essen, Essen, D-45147, Germany.
*Correspondence address. Tamas Spisak, Hufelandstraße 55 D-45147 Essen, Germany. E-mail: tamas.spisak@uk-essen.de

Abstract

Background: The lack of nonparametric statistical tests for confounding bias significantly hampers the development of robust, valid, and generalizable predictive models in many fields of research. Here I propose the *partial confounder test*, which, for a given confounder variable, probes the null hypotheses of the model being *unconfounded*.

Results: The test provides a strict control for type I errors and high statistical power, even for nonnormally and nonlinearly dependent predictions, often seen in machine learning. Applying the proposed test on models trained on large-scale functional brain connectivity data ($N = 1{,}865$) (i) reveals previously unreported confounders and (ii) shows that state-of-the-art confound mitigation approaches may fail preventing confounder bias in several cases.

Conclusions: The proposed test (implemented in the package *mlconfound*; https://mlconfound.readthedocs.io) can aid the assessment and improvement of the generalizability and validity of predictive models and, thereby, fosters the development of clinically useful machine learning biomarkers.

Keywords: machine learning, predictive modeling, confounding bias, confounder test, conditional independence, conditional permutation

- The lack of statistical tests for confounding bias hampers the development of machine learning–based biomarker candidates.
- The partial confounder test provides a model-agnostic approach for quantifying confounding bias.
- It provides strict control for type I errors and high statistical power with minimal assumptions.
- Deploying the test on functional brain connectivity data reveals that confounding bias can be problematic even if confound mitigation approaches are used.
- The test provides objective criteria to assess the specificity, generalizability, and biomedical validity of biomarker candidates.

Background

Predictive modelling uses multivariate statistical learning to aggregate information from a set of features with the aim of predicting an unknown outcome. This approach has recently become increasingly important in biomedical research and holds promise for delivering biomarkers that substantially impact clinical practice and public health [1–4]. When evaluating the usefulness and applicability of such markers, predictive performance is far from being the only important consideration. Biomedical validity and generalizability across contexts and populations are also fundamental requirements for candidate biomarkers [5–7].

Spurious, out-of-interest associations between the predictor variables (features) and the prediction target can be detrimental to the model's biomedical validity and generalizability. This phenomenon is often called confounding bias [8]. Confounding bias can be driven by various sources. For instance, measurement artifacts (e.g., motion artifacts in magnetic resonance imaging–based predictive models) are well known as a potential confounder that can bias the predictive model's output in, among others, Alzheimer's disease [9], attention-deficit/hyperactivity disorder [10, 11], or autism spectrum disorder (ASD) [12–14]). Confounding bias is, however, not restricted to measurement artifacts. Depending on the research question, several demographic and psychometric variables or the time of day of the data acquisition [15] can emerge as confounders. As a characteristic example, models trained to predict intelligence [16, 17] might provide a statistically significant predictive performance by picking up solely on age-related variance [18, 19]. Moreover, various types of systematic sampling bias, as well as stochastic group differences in the training sample, can result in confounded models (e.g., racially biased machine learning models [6, 20, 21]).

Confounding bias is especially problematic in population neuroscience studies. While large-scale multisite studies are of key importance for developing robust machine learning markers [22], most of the confounding effects are much more likely to occur in such big, longer-term studies [23], and batch and center effects may arise as additional sources of confounding bias [24, 25].

While various data-cleaning methods and dedicated prediction algorithms may help in mitigating confounding bias [9, 13, 26–29], effects of confounders can potentially bleed through into predic-

Received: April 25, 2022. **Revised:** July 7, 2022. **Accepted:** July 28, 2022

tions even if they are being attempted to control for in the prediction algorithm (see Supplementary Analysis 1 for an example), and it is often unclear which variables should be considered as confounders. In a number of cases, removing or controlling for a confounder can remove variance of interest and complicate model interpretations [24, 29, 30], rendering the choice of confound mitigation strategy as one of the most difficult compromises in predictive model development.

Powerful and robust statistical tests for quantifying confounding bias in predictive models could substantially foster both the identification of confounders to correct for and the assessment of the effectiveness of various confound mitigation approaches. It is tempting to think about confounding bias as the *conditional dependence* of the model output on the observed confounder, given the target variable. However, the proper evaluation of conditional independence among these variables is challenging. Namely, even in the presence of a slight nonnormality and/or nonlinearity of the involved conditional distributions, the "conditional" analogs of the most popular bivariate nonparametric tests (like the partial Spearman correlation; see Fig. 3) are not valid measures of conditional independence. Although warnings about this issue were given from early on [31] and received a fair amount of attention recently [32–36], the magnitude of the problem may not be fully appreciated in case of predictive model diagnostics, where nonnormality and nonlinearity of the model output can be frequently seen (see Supplementary Figs. S10–S11), as a consequence of, for example, feature-set characteristics and model regularization [37, 38].

Recently, 2 different approaches were proposed for quantifying confounding bias [39, 40]. However, these methods either fail to control type I error (as known in the case of balanced permutations [41, 42], used in Neto et al. [39]) or do not provide *P* values at all [40]. Moreover, without some modifications, they are only applicable for categorical variables and involve refitting the model, which may not be feasible for models with high computational cost (e.g., when trained with nested cross-validation).

This work aims to construct a statistical test for confounding bias that (i) guarantees valid type I error control for arbitrary models, even if nonnormal and/or nonlinear dependencies are involved; (ii) does not require refitting the model; and (iii) is applicable for classification as well as for prediction problems and both with numerical and categorical confounders.

Methods

Notation and background

In a predictive modeling setting, let $\mathbf{y}$ denote the target variable, $\mathbf{X}$ denote the feature variables, $\hat{\mathbf{y}}$ denote model output (i.e., the predictions for $\mathbf{y}$), and $\mathbf{c}$ denote a variable that is considered a confounder. Note that $\mathbf{y}$ and $\mathbf{c}$ must be observed during the experiment, whereas $\hat{\mathbf{y}}$ is provided by the predictive model. Confounding bias typically emerges in situations where $\mathbf{X} \leftarrow \mathbf{c} \rightarrow \mathbf{y}$ (arrows denoting dependence of $\mathbf{X}$ and $\mathbf{y}$ on $\mathbf{c}$), although $\mathbf{c} \rightarrow \mathbf{y}$ is not a prerequisite. After fitting the predictive model, we aim to construct predictions based on features unseen during the model training procedure: $\mathbf{X} \rightarrow \hat{\mathbf{y}}$ so that $\mathbf{y} \rightarrow \hat{\mathbf{y}}$. Obviously, a strong association between $\hat{\mathbf{y}}$ and $\mathbf{c}$ may indicate that the model is biased; its predictions are driven by the confounder rather than information about the target variable. Assessing the simple bivariate (unconditioned) dependence ($H0 : \hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c}$) between $\hat{\mathbf{y}}$ and $\mathbf{c}$ (or any of the $\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}$ variables) is, however, insufficient for the proper characterization of confounding bias in predictive modeling. For instance, even if $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c}$ is false, $\hat{\mathbf{y}}$ might be only marginally dependent on $\mathbf{c}$, due to the dependence of both on $\mathbf{y}$. In other words, if the target variable $\mathbf{y}$ displays a true association to the confounder variable $\mathbf{c}$, a model that is completely blind to $\mathbf{c}$ (i.e., not confounded at all) might still provide outputs $\hat{\mathbf{y}}$ that are significantly associated with $\mathbf{c}$.

Conditional independence for testing confounding bias

Instead of focusing on the "unconditioned" independence between the confounder and the predictions, we shall consider the *conditional independence* between $\hat{\mathbf{y}}$ and $\mathbf{c}$ given $\mathbf{y}$ (written as $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c}|\mathbf{y}$), which, by definition [43], means that $\mathbb{P}(\hat{\mathbf{y}}, \mathbf{c}|\mathbf{y}) = \mathbb{P}(\hat{\mathbf{y}}|\mathbf{y})\mathbb{P}(\mathbf{c}|\mathbf{y})$. Testing whether $\mathbf{c}$ is independent from $\hat{\mathbf{y}}$, conditional on $\mathbf{y}$, is essentially checking whether the path $\mathbf{c} \rightarrow \mathbf{X} \rightarrow \hat{\mathbf{y}}$ has been blocked in the prediction algorithm. The statistical test with the null hypothesis $H0 : \hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c}|\mathbf{y}$ will be referred to as the *partial confounder test*. Of note, although typically less useful in a predictive modeling context, one might also be interested in testing $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{y}|\mathbf{c}$. We refer to the corresponding test as the *full confounder test*.

Conditional independence—in its general form—is a fundamental concept in statistics with numerous biomedical applications [33, 34, 44, 45]. Recently, Shah and Peters [35] have raised important concerns regarding conditional independence testing. Their "no free lunch" theorem implies that, without placing some assumptions on the joint distribution of ($\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}$), conditional independence testing is effectively impossible. In other words, neither the full nor the partial confounder tests can be constructed so that—for all distributions—they provide a valid type I error control and, at the same time, a nontrivial statistical power.

This result stands in strong contrast to *unconditional* independence testing—where permutation tests [46, 47] provide a general, distribution-free solution—and it has important implications for confounder testing in predictive modeling where the distribution of the model outputs (conditioned on the target variable)—depending on the applied machine learning model—is unknown and often nonnormal and nonlinear. One of the trivial candidates for the task, partial correlation, for instance assumes that all involved variables are multivariate Gaussian and—as to be shown below in a simulated example—even its Spearman-based variant is unable to tolerate relatively small deviations from normality and linearity.

Recently, Candès et al. [33] and, based on their work, Berrett et al. [36] have demonstrated that valid and powerful conditional independence tests can be constructed with inputting distributional information about only 2 (out of the 3) variables. Specifically, the conditional permutation test (CPT) of Berrett and colleagues [36] samples from a nonuniform distribution over the set of possible permutations π of one of the variables, based on its conditional distribution of the other variable. Thereby, it incorporates the information available about the conditional distribution of interest into the permutation-based inference in a statistically valid manner.

Like many related papers, the work of Berrett et al. [36] was formalized as a (semi)supervised learning approach, where $\mathbf{X}$ is a set of predictors (features), $\mathbf{y}$ is the target variable, and $\mathbf{c}$ is a potential confounder (Fig. 1A). In this setting, testing the null hypothesis $\mathbf{X} \perp\!\!\!\perp \mathbf{y}|\mathbf{c}$ aims to determine whether the features $\mathbf{X}$ still affect $\mathbf{y}$, when controlling for $\mathbf{c}$. For instance, in genome-wide association studies, CPT can be used to determine whether a particular genetic variant $\mathbf{X}$ affects a response $\mathbf{y}$ such as disease status or some other phenotype, even after controlling for the rest of the genome, encoded in $\mathbf{c}$.

In this article, a different setting is considered, where the supervised learning model is already fitted (Fig. 1B) and we are focusing on model diagnostics by testing the triplet ($\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}$), with the

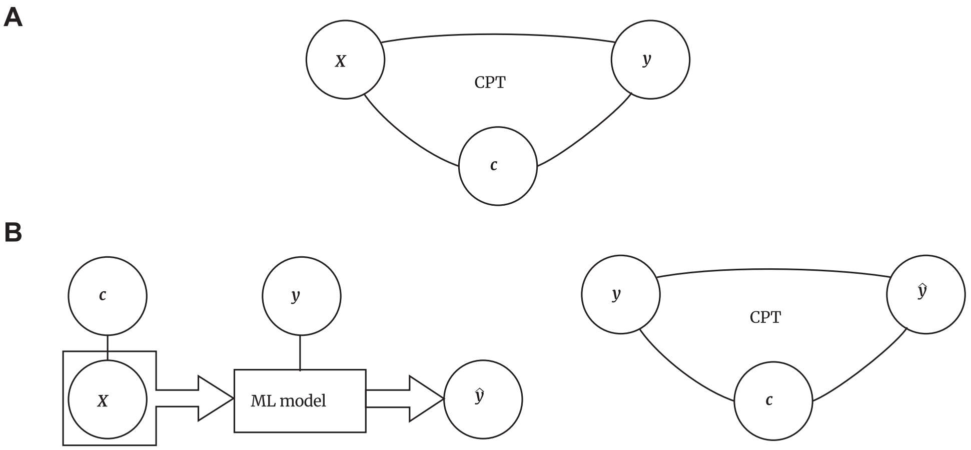


Figure 1: Conditional permutation testing as a tool for predictive model diagnostics. (A) Conditional permutation testing (CPT) was originally proposed to be used on the feature variable X, target variable Y, and confounders Z, to perform statistical inference. (B) The proposed use of CPT in predictive modeling requires the model to be fitted first, to obtain the model's prediction $\hat{\mathbf{y}}$ on $\mathbf{y}$. CPT is then utilized on the triplet $(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c})$, to test hypotheses $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c} | \mathbf{y}$ or $\mathbf{y} \perp\!\!\!\perp \hat{\mathbf{y}} | \mathbf{c}$. Using CPT this way allows lifting assumptions on the prediction target. However, as shown in Fig. 3, the original can still provide inflated P values in case of nonlinearity in the conditional distributions. False positives can be successfully eliminated by the proposed nonlinear techniques for conditional distribution modeling (Fig. 2).

Table 1. Possibilities when testing conditional independence in potentially biased predictive models. The table lists the 3 possible null hypotheses (H0) and the variables where assumption about the joint/conditional distributions is required/not required ($\mathbf{y}$: prediction target, $\hat{\mathbf{y}}$: predictions, $\mathbf{c}$: confounder variable).

		H0	Assumption needed for	No assumptions about the distribution of
1.	$\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{y} \| \mathbf{c}$	Full confounder test: model exclusively driven by the confounder	$Q(\mathbf{y}\|\mathbf{c})$	$(\hat{\mathbf{y}}, \mathbf{y})$, $(\hat{\mathbf{y}}, \mathbf{c})$
2.	$\mathbf{y} \perp\!\!\!\perp \mathbf{c} \| \hat{\mathbf{y}}$	Model captures all variance in the confounder (not of interest)	$Q(\mathbf{c}\|\hat{\mathbf{y}})$	$(\mathbf{y}, \mathbf{c})$, $(\mathbf{y}, \hat{\mathbf{y}})$
3.	$\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c} \| \mathbf{y}$	Partial confounder test: model not directly driven by the confounder	$Q(\mathbf{c}\|\mathbf{y})$	$(\hat{\mathbf{y}}, \mathbf{c})$, $(\hat{\mathbf{y}}, \mathbf{y})$

requirement of minimal assumptions on the conditional distribution of $\hat{\mathbf{y}}$ on $\mathbf{y}$ and $\mathbf{c}$ (Fig. 1C).

Within this setting, conditional independence testing and, specifically, the framework of conditional permutation testing allows investigating 3 different null hypotheses corresponding to the $(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c})$ triplet. As listed in Table 1, testing the null hypothesis $\mathbf{y} \perp\!\!\!\perp \hat{\mathbf{y}} | \mathbf{c}$ (option 1, full confounder testing) investigates whether the predictions are likely explainable solely with the confounder (i.e., whether the model is exclusively confounder driven). Testing $\mathbf{y} \perp\!\!\!\perp \mathbf{c} | \hat{\mathbf{y}}$ (option 2) addresses whether the model captures all the variance in c when predicting y. Testing the null hypothesis $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c} | \mathbf{y}$ (option 3, partial confounder testing) examines whether the dependence of the model output on the confounder can likely be explained by the confounder's dependence on the target variable (i.e., whether there is any confounding bias in the model).

Option 3 (i.e., partial confounder testing) is typically of interest when testing confounding bias of predictive models. Option 1 (i.e., full confounder testing) may be less useful in practice, although it might provide valuable insights in the exploratory phase of model construction. Option 2 does not seem appealing for model diagnostics, and importantly, in this case, the proposed variety of the CPT framework does not allow constructing a test that is nonparametric on $\hat{\mathbf{y}}$. We will therefore focus on option 3 (i.e., the partial confounder test).

In the following section, CPT is adapted for *partial* confounder testing and extended with the general additive model [48] (GAM) and multinomial logistic regression [49, 50] based conditional distribution estimations, in order to make it handle categorical data and nonlinear dependencies between the confounder and the target variable. (For an overview of the method, see Fig. 2.)

The partial confounder test

The inner workings of the *partial confounder test* are summarized in Fig. 2. In short, the test models the conditional distribution between the confounder and the target variable with a GAM—or with an *mnlogit* regression, in case of a categorical confounder—and then uses a so-called parallel-pairwise Markov chain Monte Carlo sampler of Berrett et al. [36] that draws permutations of the original confounder, so that the permuted variables still comply with the estimated conditional distribution. As a result, the permuted "copies" of the confounder variable retain its correlation with the target variable but eliminate any "additional" relationship with the model output. The test statistic (coefficient of determination, R^2) is then computed between the model output and the original, as well as the permuted confounder variables. The original and the permuted test statistics construct the P value as the ratio of permuted test statistics more extreme than the original.

In detail, the partial confounder test generates a null distribution for an arbitrary predefined test statistic $T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c})$ by sampling permutation based "copies" of the original $\mathbf{c}$,

$$c_i^{(j)} \sim Q(\cdot|y_i) \qquad (1)$$

where, $Q(.|y)$ denotes the conditional distribution of $\mathbf{c}$ given $\mathbf{y} = y_i$ and $j = 1, \ldots, m$ indexes the "copy" of $\mathbf{c}$ so that

$$\mathbf{c}^{(j)} = (c_1^{(j)}, \ldots, c_n^{(j)}) = (c_{\pi_1^{(j)}}, \ldots, c_{\pi_n^{(j)}}) = \mathbf{c}_{\boldsymbol{\pi}^{(j)}}$$

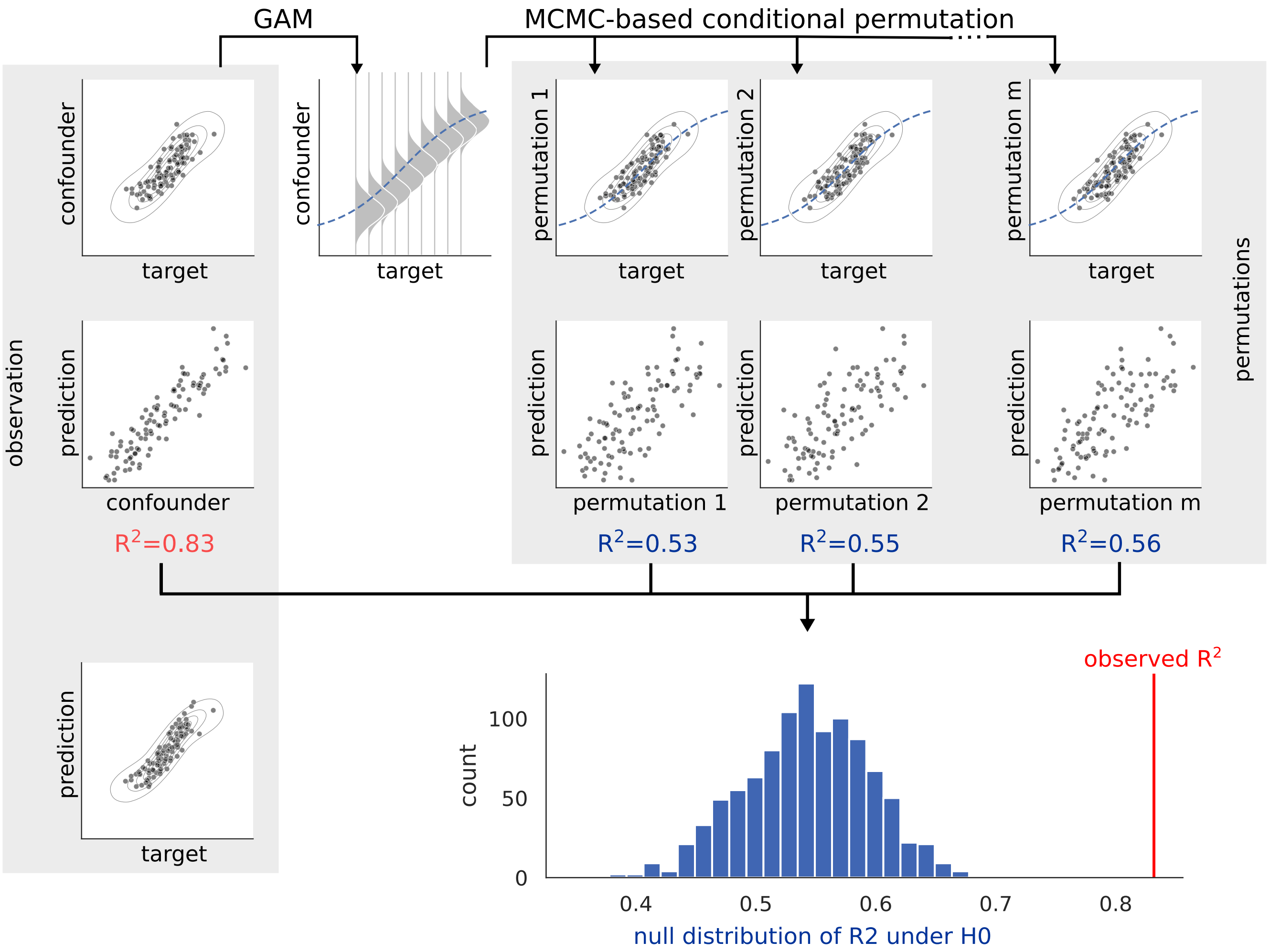


Figure 2: Graphical representation of the proposed partial confounder test. The partial confounder test models the conditional distribution of the confounder, given the target variable, with a generalized additive model (GAM). The parallel-pairwise Markov chain Monte Carlo (MCMC) sampler draws permutations of the original confounder variable that comply with the GAM-based conditional distribution (permutation 1, 2,..., m). The test statistic (coefficient of determination, R^2) is then computed between the model output and the original, as well as the permuted confounder variables. The original and the permuted test statistics construct the P value as the ratio of permuted test statistics more extreme than the original. Figure source code available as jupyter notebook: https://github.com/pni-lab/mlconfound-manuscript/blob/main/simulated/overview-fig.ipynb.

is a permutation of the original vector $\mathbf{c} = (c_1, \ldots, c_n)$, with its elements reordered according to the permutation $\boldsymbol{\pi} \in S_n$, where S_n denote the set of all permutations on the indices $\{1, \ldots, n\}$.

As shown by Berrett et al. [36], to ensure that Equation 1 holds, the $\mathbf{c}_{\boldsymbol{\pi}^{(j)}}$ copies must be drawn so that

$$\mathbb{P}(\boldsymbol{\pi}^{(j)} = \boldsymbol{\pi}|\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}) = \frac{q^n(\mathbf{c}_{\boldsymbol{\pi}}|\mathbf{y})}{\sum_{\boldsymbol{\pi}' \in S_n} q^n(\mathbf{c}_{\boldsymbol{\pi}'}|\mathbf{y})} \tag{2}$$

that is, according to the $q^n(\cdot|\mathbf{y}) := q(\cdot|y_1) \ldots q(\cdot|y_n)$ product density corresponding to the conditional distribution $Q(\cdot|\mathbf{y})$. Note that Equation 2 does not necessarily assume a continuous distribution.

This mechanism creates copies $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(m)}$ so that under the null hypothesis $(\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{c}|\mathbf{y})$, the triples

$$(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}), (\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(1)}), \ldots, (\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(m)})$$

are all identically distributed and so are the

$$T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}), T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(1)}), \ldots, T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(m)})$$

test statistics, as well.

As long as the numerator of Equation 2 is nonzero for all $c_\pi \in C$ and $y \in Y$, the conditional permutations constitute an algebraic group; thus, as shown by Hemerik and Goeman [42], an unbiased estimate of the P value under the null can be obtained as

$$p = \frac{\sum_{j=1}^{m} \mathbb{1}\{T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(j)}) \geq T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c})\}}{m}$$

While the group property of the conditioned permutations provides a straightforward proof for the validity of the approach, for an alternative verification, see the proof of Theorem 1 in Berrett et al. [36].

The required permutations could be theoretically sampled with a simple Metropolis–Hastings algorithm that draws uniformly from S_n at random. However, this way, the acceptance ratio would be extremely low, even for moderate n (except there is very low dependence of $\mathbf{c}$ on $\mathbf{y}$), resulting in slow mixing times. The partial confounder test can be, however, efficiently implemented with the parallelized pairwise Markov chain Monte Carlo sampler of Berrett et al. [36] (Algorithm 1), which draws disjoint pairs in parallel and decides whether or not to swap them randomly, according to the odds ratio calculated from the conditional densities belonging to the original and swapped data. The acceptance odds ratio of swapping indices i and j is

$$\ln \frac{q(c_j|y_i)q(c_i|y_j)}{q(c_i|y_i)q(c_j|y_j)} = \ell(c_j|y_i) + \ell(c_i|y_j) - \ell(c_i|y_i) - \ell(c_j|y_j) \quad (3)$$

where ℓ denotes the log-likelihood.

In their Theorem 2, Berrett et al. [36] verify that the resulting Markov chain yields the desired stationary distribution, even if the number of steps is small.

Conditional log-likelihood

Obtaining a relatively accurate, independent estimate of $Q(\cdot|\mathbf{y})$ (of any shape) for CPT inference is important. Berrett and colleagues [36] recommend using a large independent sample to obtain the log-likelihood matrix that represents the conditional distribution $Q(\cdot|Z)$ or, alternatively, to reuse the data by fitting a least squares linear regression:

$$\mathbf{c} = \alpha + \beta \mathbf{y} + \boldsymbol{e} \quad (4)$$

As the linear regression-based method, obviously, does not handle nonlinear relationships, I propose to apply a modeling approach that accounts for nonlinearity. Although several nonparametric techniques might be suitable for this purpose, many of these tend to be greedy for large sample sizes, may lack stability, or perform poorly with many potential predictors. Certain methods, such as kernel methods and smoothing splines, are also very difficult to interpret [51]—an important consideration when analyzing the source of a confounder effect.

Here, I propose to use the GAM of Hastie and Tibshirani [48]:

$$\mathbf{c} = \alpha + \beta f(\mathbf{y}) + \boldsymbol{e} \quad (5)$$

where the feature function f is built using penalized B-splines, which allow us to automatically model nonlinear relationships without having to manually try out many different transformations on each variable. The principal advantages of GAM are that (i) the complexity of the model can be effectively regularized trough its hyperparameters, (ii) it is able to model highly complex nonlinear relationships with a potentially large number of both numeric and categorical predictors, and (iii) it has computationally effective solver algorithms. The potential disadvantages of GAMs are not relevant for the problem at hand or can be easily overcome. Specifically, the possibly poor out-of-distribution generalization of GAM is not problematic, as in our approach, the model is not used for constructing out-of-distribution predictions. Moreover, as several other models, GAMs can easily overfit the data. However, in the proposed approach, the smoothness of the GAM model is optimized with a grid search by picking the model with the lowest generalized cross-validation score from the models defined by the default parameters as implemented in PyGAM [52] (v0.8.0).

If we write $\boldsymbol{\mu} = \alpha + \beta f(\mathbf{y})$ and σ denotes the standard deviation of the residual $\boldsymbol{e}$, then the conditional distribution of interest can be assumed to be normal with the parameters

$$(\mathbf{c}|\mathbf{y} = y_i) \sim \mathcal{N}\{\mu_i, \sigma^2\}$$

and the log-likelihood, which is to be used in Equation 3, can be computed simply as the log of the corresponding probability density function:

$$\ell(c_i|y_j) = -\frac{1}{2}\left(\frac{c_i - \mu_j}{\sigma}\right)^2 - ln(2\pi\sigma)$$

In the case of categorical $\mathbf{c}$, a multinomial logistic regression (*mnlogit*) model can be used to obtain $D(\cdot|\mathbf{y})$, with the extra assumption of *complete separation* if $\mathbf{y}$ is also categorical (in order to ensure an invertible Hessian, see, e.g., [49, 50]).

Importantly, both the GAM- and the *mnlogit*-based approaches guarantee that the numerator of Equation 2 is always greater than zero and the group property for the permutations holds.

Note that from the 3 options for conditional independence-based null hypotheses enumerated in Table 1, the proposed approach cannot provide a test for option 2 that is assumption free about $\hat{\mathbf{y}}$, as the variable, on which the independence is conditional, must be always the predictor variable in Equation 5. However, as discussed above, this option is of low practical relevance anyway. Pleasingly, the proposed Gaussian regression-based conditional likelihood estimation ensures that no assumptions on $\hat{\mathbf{y}}$ have to be made for the practically relevant options 1 and 3 (i.e., for the full and partial confounder tests).

In theory, any predefined test statistic T can be used with the proposed approach. The Python package *mlconfound*, implementing the proposed full and partial confounder tests, utilizes the coefficient of determination (R^2 or pseudo-R^2 in case of categorical confounder or classification [53]) as a test statistic: $T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}) = R^2(\hat{\mathbf{y}}, \mathbf{c})$ and $T(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{c}^{(j)}) = R^2(\hat{\mathbf{y}}, \mathbf{c}^{(j)})$, which allows interpretable, 2-tailed inference.

Validation on simulated data

Using CPT to test confounding bias in predictive modeling allows relaxing assumptions on $\hat{\mathbf{y}}$ but—in line with the "no free lunch" theorem, requires knowing—or putting assumptions on—the joint distribution of the other 2 variables ($\mathbf{y}$ and $\mathbf{c}$). Berrett et al. [36] give a detailed analysis of the robustness of their CPT approach when estimating the conditional distribution with reusing the tested data via linear regression and, also, against misspecifying the conditional distribution of interest to introduce nonlinearity.

Here I extend these results by performing simulations that evaluate the GAM- and *mnlogit*-based approaches, in a form that is accessible for power calculations in predictive modeling (considering various weights of the target signal in $\mathbf{c}$ and the confounder and the target signals in $\hat{\mathbf{y}}$). Moreover, I investigate the robustness of the tests against the violation of normality and linearity of the conditional distributions $D(\mathbf{c}|\mathbf{y})$ and $D(\hat{\mathbf{y}}|\mathbf{y})$.

Simulations are performed separately for the 2 proposed tests.

Simulation approach

As a first step, the target variable $\mathbf{y}$ is drawn randomly from a normal distribution:

$$\mathbf{y} \sim \mathcal{N}(0, 1)$$

Next, the confounder signal is simulated as

$$\mathbf{c}|y_i \sim f_{\delta,\epsilon}(\mathcal{N}(0, 1)) + w_{yc}\, g(y_i)$$

where f is a function to introduce nonnormality, namely, the *sinh-arcsinh* transformation of Jones and Pewsey [54], defined as

$$f_{\delta,\epsilon}(\mathbf{x}) = sinh(\delta sinh^{-1}(\mathbf{x}) - \epsilon)$$

where the parameters δ and ϵ control the kurtosis and skewness of the resulting *sinh-arcsinh* distribution, with $\delta = 1$ and $\epsilon = 0$ producing the identity function (i.e., no nonnormality introduced).

Moreover, nonlinearity can be introduced with the function g, which can be simply the identity function (no nonlinearity is introduced in this case) or, for instance, a sigmoid-shaped function, in our case:

$$g(\mathbf{x}) = tanh(\mathbf{x})$$

The simulated predicted values are constructed in a similar fashion but may depend on $\mathbf{c}$ as well:

$$\hat{\mathbf{y}}|y_i, c_i \sim f_{\delta,\epsilon}(\mathcal{N}(0,1)) + w_{y\hat{y}}\, g(y_i) + w_{c\hat{y}}\, c_i$$

Note that simulations with $w_{c\hat{y}} = 0$ produce data under the null hypothesis of no confounding bias.

To test the implementation for categorical variables, simulated $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\mathbf{c}$ variables are binarized by thresholding at 0.

Simulations for comparison with partial Spearman correlation and linear CPT

To demonstrate the need for the proposed GAM-based CPT approach for partial confounder testing (Fig. 3), its validity was contrasted to partial Spearman correlation and the linear variety of CPT (based on Equation 4, as described by Berrett et al. [36]) with the following simulation parameters: sample size $n = 1{,}000$, $w_{c\hat{y}} = 0$ (i.e., H0 simulations only), taking all combinations of $w_{yc} \in \{0.5, 1, 2, 3\}$ and $w_{y\hat{y}} \in \{0.5, 1, 2, 3\}$. Furthermore, simulations cases with nonnormality ($f_{\delta=0.1,\epsilon=2}$) and nonlinearity (sigmoid g) have also been investigated for all simulation cases.

For each parameter combination, 1,000 repetitions were performed and false-positive rates were calculated as the ratio of P values smaller than $\alpha = 0.05$.

The simulation cases are exemplified (with $w_{yc} = w_{y\hat{y}} = 2$) on the left of Fig. 3.

Simulations for evaluating power

One hundred repetitions were performed of all combinations of the following parameter values: $w_{yc} \in \{0.5, 1, 2, 3\}$, $w_{y\hat{y}} \in \{0.5, 1, 2, 3\}$, $w_{c\hat{y}} \in \{0, 0.2, 0.4, 0.6\}$, $n \in \{50, 100, 500, 1{,}000\}$. All simulations were performed with both linear and sigmoid dependence as well as with normal and nonnormal conditional distributions: $(\delta, \epsilon) = \{(0.1, 2), (1, 0), (1.05, -3), (1.5, -5), (5, -10)\}$.

The partial confounder tests, as implemented in version 0.20.0 of the package "*mlconfound*," were run with default parameters (1,000 permutations and 50 Markov chain Monte Carlo steps to generate the conditioned permutations) and by implying categorical variables, where needed.

All code used for the simulations is available on GitHub (https://github.com/pni-lab/mlconfound-manuscript/tree/main/simulated).

Application on functional brain connectivity data

The usefulness of the proposed confounder tests is demonstrated by applying them for predictive classification and regression models based on functional brain connectivity data, processed with different confound mitigation approaches.

Partial confounder testing was performed with 10,000 permutations and 50 Markov chain Monte Carlo steps, as implemented in version 0.20.0 of the package "*mlconfound*." Unconditional dependence across the involved variables was investigated with conventional permutation tests on the R^2 values, with 1,000 permutations.

All empirical analyses are available as jupyter notebooks on GitHub (https://github.com/pni-lab/mlconfound-manuscript/tree/main/empirical).

HCP: testing age and acquisition batch bias in fluid intelligence prediction

The Human Connectome Project dataset contains imaging and behavioral data of approximately 1,200 healthy subjects [55]. Preprocessed resting state functional magnetic resonance imaging (fMRI) connectivity data (partial correlation matrices) [56] as published with the HCP1200 release ($N = 999$ participants with functional connectivity data) were used to build models that predict individual fluid intelligence scores (G_f), measured with Penn Progressive Matrices [57].

To ensure normality of the target variable for the partial correlation-based analyses, G_f was nonlinearly transformed to normal distribution with the quantile transformation [58] as implemented in *scikit-learn* [59] (see Supplementary Fig. S8 for details).

Features (functional connectivities across 100 group-independent component analysis–based regions) were either (i) considered in their raw form or were subject to confound mitigation approaches by (ii) feature regression [9] or (iii) COMBAT [28, 60]. The feature mitigation strategies were separately applied for acquisition batch and age group as confounder variable.

Each of the 5 types of features (raw, regressing out acquisition batch, regressing out age group, COMBAT with acquisition batch, COMBAT with age group) was independently incorporated into a *scikit-learn*–based [59] machine learning procedure aiming to predict the individual fluid intelligence scores with a ridge regression [61]. The α parameter of the ridge model was considered a hyperparameter ($\alpha \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000\}$) and optimized in a nested cross-validation with 10 folds in both the inner and the outer loops and with mean squared error as an optimization metric. Confound mitigation was performed inside of the outer cross-validation loop, to avoid leakage.

ABIDE: testing motion and center bias in predictive models of ASD diagnosis

The proposed tests were applied to provide evidence of center and motion bias in diagnostic predictive models of ASD, trained on the Autism Brain Imaging Data Exchange (ABIDE) dataset [62] involving 866 participants (ASD: 402, neurotypical control: 464). Preprocessed regional time-series data were obtained as shared (https://osf.io/hc4md) by Dadi et al. [63], which were based on preprocessed image data provided by the Preprocessed Connectome Project [64].

Tangent correlation across the time series of the $n = 122$ regions of the BASC (Multi-level bootstrap analysis of stable clusters) [65] brain parcellation was computed with nilearn (http://nilearn.github.io/) [66, 67].

The resulting functional connectivity estimates were considered features either (i) in their raw form or after applying (ii) feature regression [9] or (iii) COMBAT [28, 60]. The investigated confounder variables were "imaging center" and "in-scanner motion," as measured by the mean framewise displacement (FD), as defined by Power et al. [68]. Mean FD was nonlinearly transformed to normal distribution with the quantile transformation [58] as implemented in *scikit-learn* [59] (see Supplementary Fig. S9 for details).

As COMBAT is not able to handle continuous variables (since it was primarily designed to remove categorical "batch effects"), motion was binned into 10 groups, based on equidistant data quantiles ranging from 0 to 1.

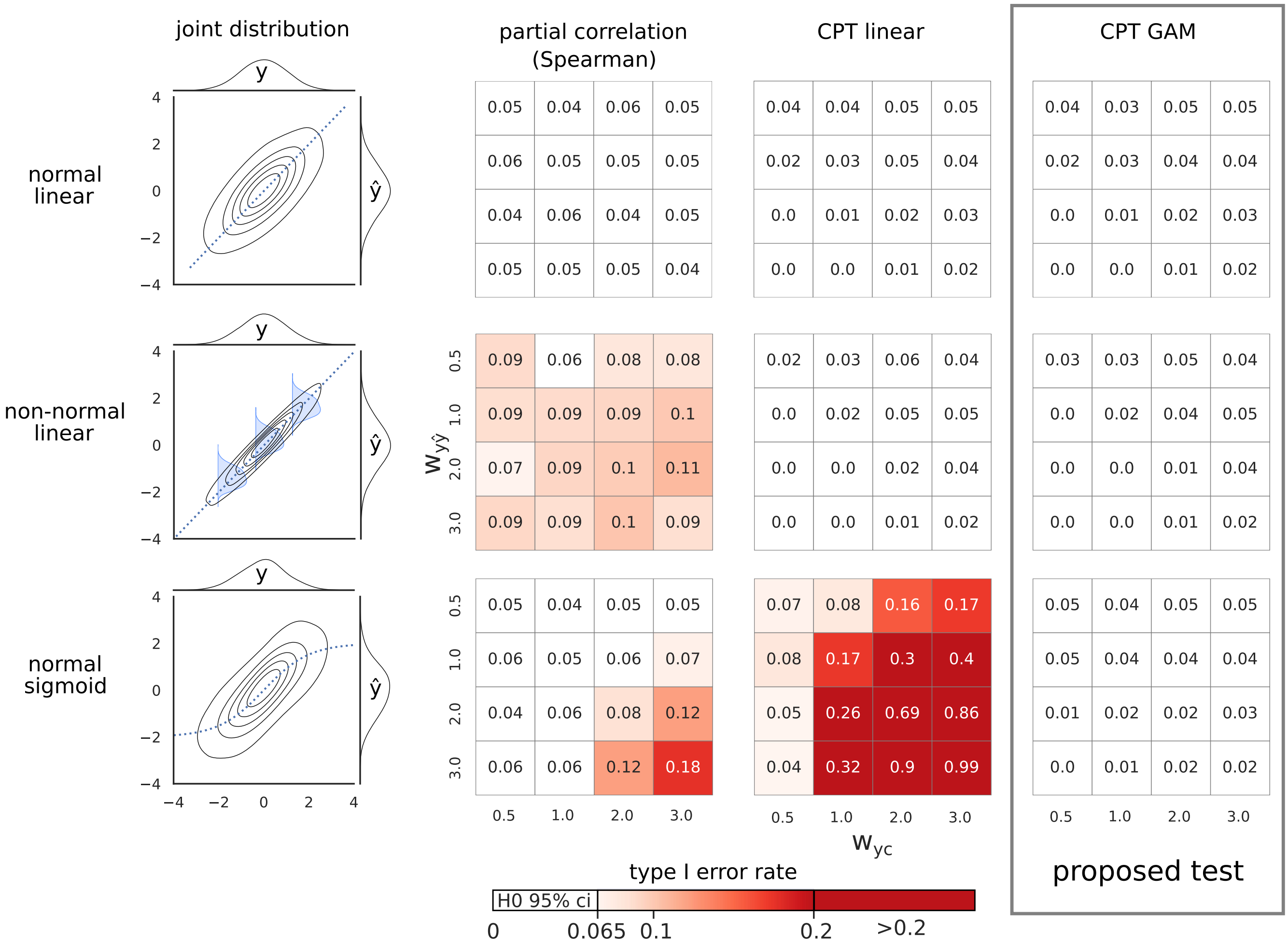


Figure 3: Type I error control of partial Spearman correlation, linear and GAM-based conditional permutation test. Type I error control was investigated in 3 example cases: normal conditional distribution with linear dependency (first row), slightly nonnormal conditional distribution with linear dependency (second row), and normal conditional distribution with nonnormal (sigmoid) dependency (third row). Nonnormal conditional distribution on the second plot is illustrated with blue density diagrams (kurtosis: –0.8, skewness: –0.1). False-positive rates for confounder contributions (w_{yc}, ranging from 0.5 to 3.0) and predictive performances ($w_{y\hat{y}}$, ranging from 0.5 to 3.0) are shown in heatmaps. The upper limit for the binomial confidence interval corresponding to alpha = 0.05 is 0.065. Values below this threshold (colored white) indicate a valid type I error control.

A total of 5 types (raw, feature regression of site, feature regression of motion, COMBAT with site, COMBAT with motion) of features were independently incorporated into a *scikit-learn*–based [59] machine learning procedure aiming to predict the diagnosis (DX: ASD vs. neurotypical controls) with an L2-regularized logistic regression, as previously recommended [63]. The C parameter of the model was considered a hyperparameter ($C \in \{0.1, 1, 10\}$) and optimized in a nested cross-validation with 10 folds both in the inner and the outer cross-validation loop and with area under the receiver operator curve (AUC under ROC) as the optimization metric. Confound mitigation was performed inside of the outer cross-validation loop, to avoid leakage. Confounder testing was performed on the predicted class probabilities.

Results

Partial confounder tests

The proposed *partial confounder tests* have been implemented in the Python package *mlconfound* (https://mlconfound.readthedocs.io) (biotools:mlconfound, RRID:SCR_022545).

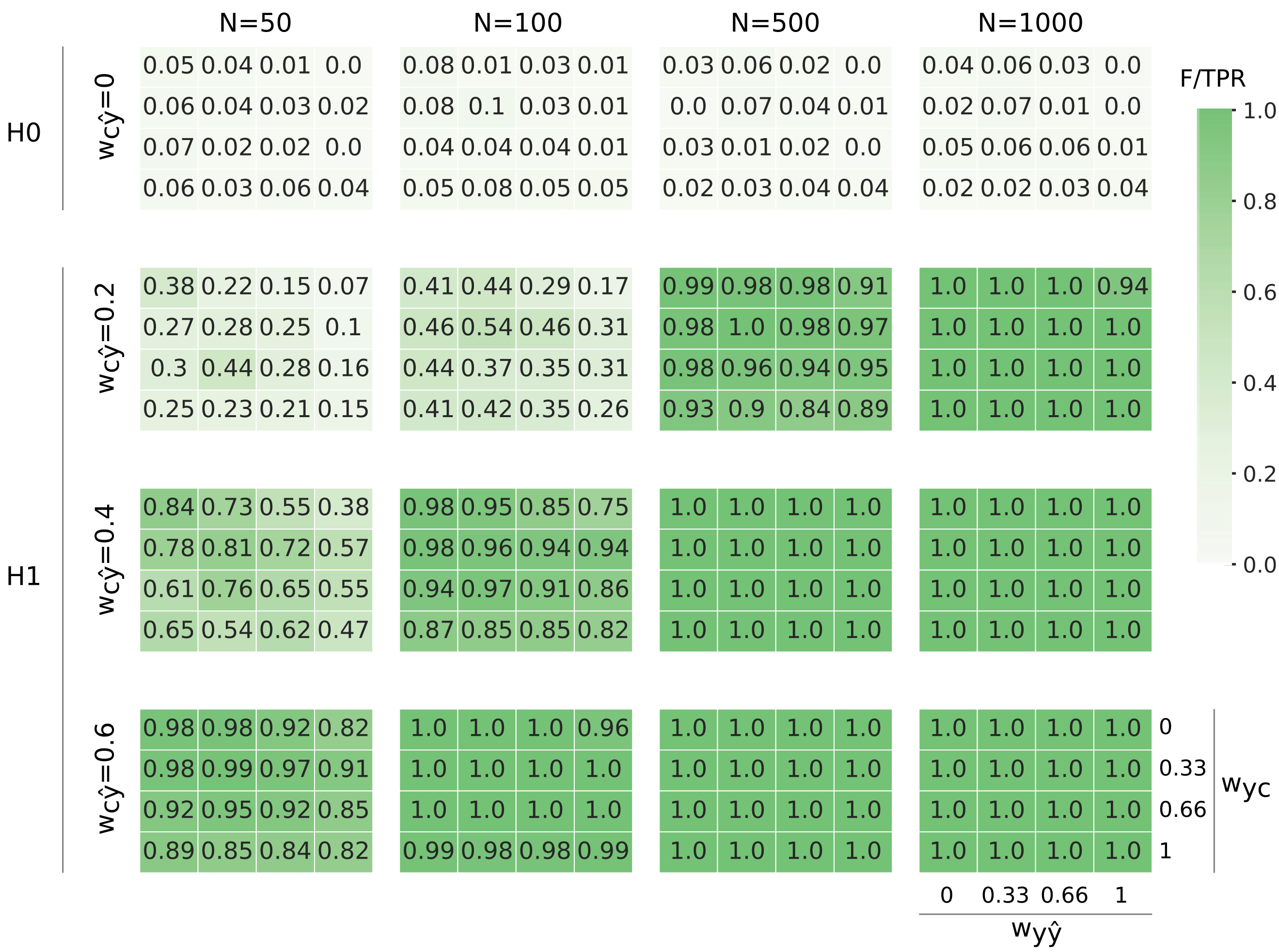
Simulations

Type I error

As suggested by theory (see Methods for details) and shown by the simulations with a wide range of settings, both of the proposed tests provide a valid type I error control (Fig. 4 and Supplementary Figs. S1–S3), even in case of nonlinearity and nonnormality (Figs. 3, 5 and Supplementary Figs. S4–S7), except when nonnormality is extreme (purple distribution on Fig. 5, kurtosis: 42, skewness: –6).

Power

Estimates of statistical power for the partial confounder test (with normal and linear simulations, for a wide range of parameters) were found to be virtually identical to those of Pearson's partial correlation (see Fig. 4 and Supplementary Fig. S12). Notably, with sample sizes as large as 1,000, a confounder contributing only ∼ 4% to the variance of the predictions ($w_{c\hat{y}} = 0.2$) can already be robustly detected with a power of 94–100%. With a sample size of 500, the same confounding bias is still detected with a power greater than 84–100% in all of the simulation cases. A sample size



		N=50				N=100				N=500				N=1000			
wcŷ	wyc \ wyŷ	0	0.33	0.66	1	0	0.33	0.66	1	0	0.33	0.66	1	0	0.33	0.66	1
wcŷ=0	0	0.05	0.04	0.01	0.0	0.08	0.01	0.03	0.01	0.03	0.06	0.02	0.0	0.04	0.06	0.03	0.0
	0.33	0.06	0.04	0.03	0.02	0.08	0.1	0.03	0.01	0.0	0.07	0.04	0.01	0.02	0.07	0.01	0.0
	0.66	0.07	0.02	0.02	0.0	0.04	0.04	0.04	0.01	0.03	0.01	0.02	0.0	0.05	0.06	0.06	0.01
	1	0.06	0.03	0.06	0.04	0.05	0.08	0.05	0.05	0.02	0.03	0.04	0.04	0.02	0.02	0.03	0.04
wcŷ=0.2	0	0.38	0.22	0.15	0.07	0.41	0.44	0.29	0.17	0.99	0.98	0.98	0.91	1.0	1.0	1.0	0.94
	0.33	0.27	0.28	0.25	0.1	0.46	0.54	0.46	0.31	0.98	1.0	0.98	0.97	1.0	1.0	1.0	1.0
	0.66	0.3	0.44	0.28	0.16	0.44	0.37	0.35	0.31	0.98	0.96	0.94	0.95	1.0	1.0	1.0	1.0
	1	0.25	0.23	0.21	0.15	0.41	0.42	0.35	0.26	0.93	0.9	0.84	0.89	1.0	1.0	1.0	1.0
wcŷ=0.4	0	0.84	0.73	0.55	0.38	0.98	0.95	0.85	0.75	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.33	0.78	0.81	0.72	0.57	0.98	0.96	0.94	0.94	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.66	0.61	0.76	0.65	0.55	0.94	0.97	0.91	0.86	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	1	0.65	0.54	0.62	0.47	0.87	0.85	0.85	0.82	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
wcŷ=0.6	0	0.98	0.98	0.92	0.82	1.0	1.0	1.0	0.96	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.33	0.98	0.99	0.97	0.91	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.66	0.92	0.95	0.92	0.85	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	1	0.89	0.85	0.84	0.82	0.99	0.98	0.98	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Figure 4 The partial confounder test provides a strict control for type I errors and a high statistical power in simulated data. Heatmaps depict positive rates (ratio of P values lower than 0.05, color coded as shown by the palette on the right) in various simulation settings (100 simulations per tile) with different simulation weights $w_{y\hat{y}}$ (predictive performance; horizontal axis on each heatmap), w_{yc} (confounder–target association; vertical axis on each heatmap), $w_{c\hat{y}}$ (degree of confounder bias; rows), and for different sample sizes (N, columns). Weights 0.2, 0.33, 0.4, 0.6, 0.66, and 1.0 can be assigned to the following approximate explained variance values: 4%, 10%, 12%, 25%, 30%, and 50%, respectively. First row contains simulations under the null hypothesis (H0, no confounding bias), and rows 2–4 represent simulations from the alternative hypothesis (H1, confounding bias). Positive rates for the simulations under the null and the alternative hypotheses can be interpreted as type I error rate and statistical power, respectively. The higher 95% confidence limit for a positive rate of alpha = 0.05 is 0.11 for each tile.

of 100 requires a somewhat stronger bias with approximately 12% of explained variance ($w_{c\hat{y}} = 0.4$) to achieve a reasonable level of power (75–98%). Finally, even with a relatively low sample size of 50, the same amount of confounder variance is detected with a power of at least 50%. If the confounder explains more than 25% of variance, it is almost certainly detected even with a low sample size of $n \geq 50$.

Simulations show that nonnormality has a minimal effect on the power of the tests, except in case of extreme nonnormality (Fig. 5). Simulations with sigmoid dependence resulted in an apparent loss of statistical power, but this is simply a consequence of the simulation methodology: with the same parameters, the sigmoid-transformed confounder explains only approximately half the variance as compared to linear simulations. Type I error control was valid in case of categorical variables, as well (Supplementary Figs. S1, S3, S5, S7).

Neuroimaging data

To demonstrate the usefulness of the proposed tests in detecting various types of confounding bias, they have been deployed in 2 typical research scenarios—a regression and a classification problem—where confounder effects are known to hamper the development of biomedically useful predictive models. The empirical analyses confirmed the presence of nonlinearity and nonnormality in the output of the predictive models (see Supplementary Fig. S11 for more details).

HCP dataset

Functional connectivity data from the Human Connectome Project [55] (HCP) were used to build predictive models of fluid intelligence (G_f) and to test for the previously discussed confounding effect of age [18, 19] and, additionally, the—somewhat underdiscussed—batch-like effect of acquisition date of the data within the course of the data acquisition process.

Both acquisition batch and age group were statistically significantly associated with G_f ($R^2 = 0.032$ and 0.011 and $P < 0.001$ and $P = 0.001$, respectively; see also Table 2). The model trained on the raw (unadjusted) connectivity features predicted fluid intelligence with a medium effect size ($R^2 = 0.095$, $P < 0.001$).

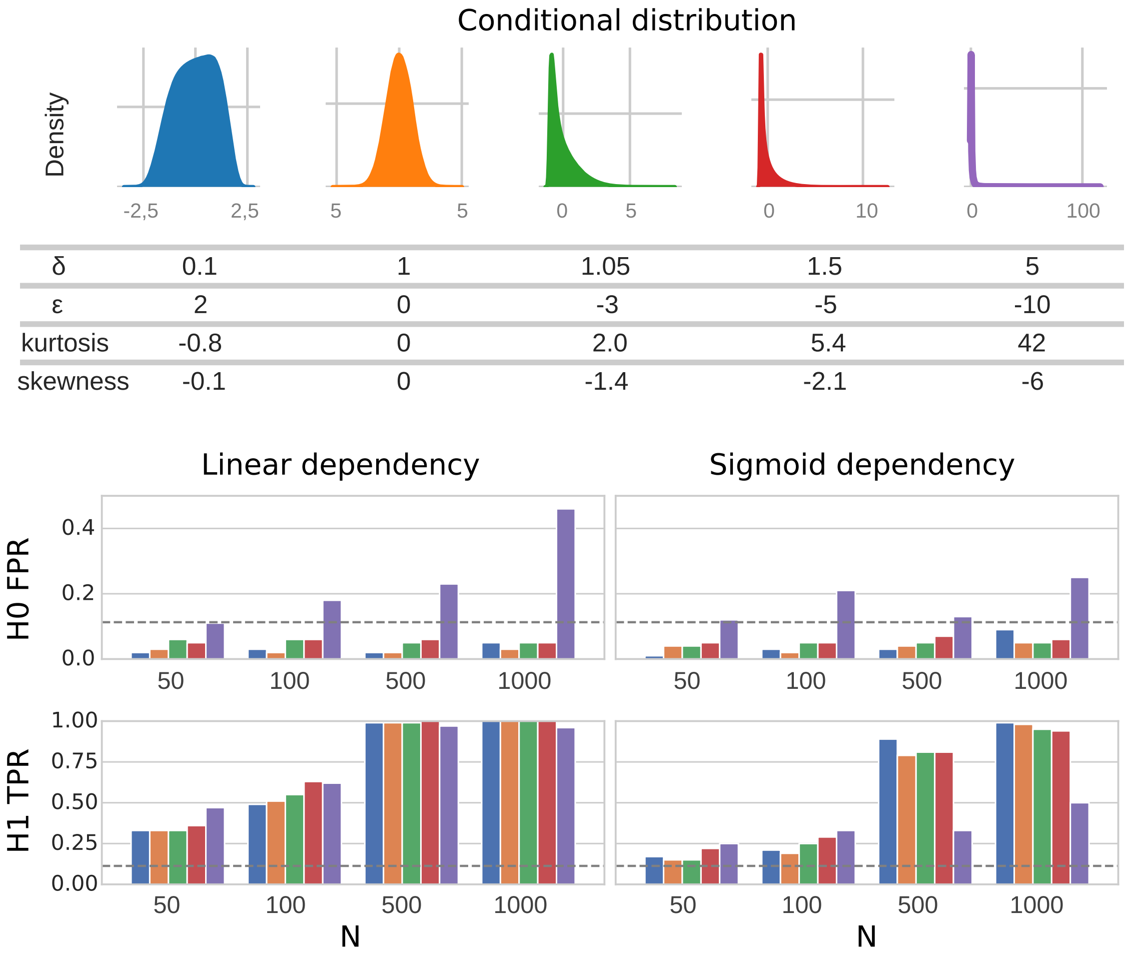


Figure 5 The partial confounder test is robust to nonnormality and nonlinearity.
Simulations included variables with 5 different degrees of nonnormality (top panel), as introduced with various δ and ϵ values of the *sinh-arcsinh* transformation (yellow: normally distributed). Fisher's kurtosis and skewness are given for each distribution. False- and true-positive rates in the simulations under H0 and H1, respectively, for each investigated sample size (N), are depicted by barplots for both linear and sigmoid dependency structure. Upper 95% binomial confidence limit corresponding to alpha = 0.05 is shown with a vertical dashed line.

Table 2. Coefficients of determination (R^2), the corresponding *P* values, and the *P* values of the partial confounder tests, for all investigated datasets, confounders (conf.), and confounder mitigation methods (method). Bold numbers denote significant confounding bias identified by the partial confounder test.

Dataset	**Conf.**	**Method**	$R^2_{y,c}$	$p_{y,c}$	$R^2_{\hat{y},c}$	$p_{\hat{y},c}$	$R^2_{\hat{y},y}$	$p_{\hat{y},y}$	**Partial confounder test**
HCP	acq.	raw	0.032	<0.001	0.071	<0.001	0.095	<0.001	**<0.0001**
		f.reg.			0.0	1.0	0.114	<0.001	1.0
		COMBAT			0.013	0.4	0.122	<0.001	0.65
	age	raw	0.011	0.001	0.034	<0.001	0.095	<0.001	**<0.0001**
		f.reg.			0.0	0.92	0.118	<0.001	0.95
		COMBAT			0.005	0.048	0.121	<0.001	0.16
ABIDE	center	raw	0.019	<0.001	0.169	<0.001	0.126	<0.001	**<0.0001**
		f.reg.			0.004	1.0	0.179	<0.001	1.0
		COMBAT			0.05	0.001	0.17	<0.001	**0.009**
	motion	raw	0.028	<0.001	0.111	<0.001	0.126	<0.001	**<0.0001**
		f.reg.			0.002	0.16	0.098	<0.001	0.51
		COMBAT			0.002	0.19	0.111	<0.001	0.59

The partial confounder test revealed that the “raw” model (without confounder mitigation) was significantly biased by both age group and acquisition batch (both $P < 0.0001$, first column of Fig. 6), with later phases of the acquisition and lower age being associated with larger predicted values.

After applying confound mitigation approaches (feature regression or COMBAT), the partial confounder test did not provide evidence for confounding bias anymore ($P > 0.05$ for all; shown in the second and third columns of Fig. 6), neither for acquisition batch nor for age. Both feature regression and COMBAT increased the predictive performance, with COMBAT providing the overall best performances ($R^2 = 0.122$ and 0.121 when applied to remove the effect of acquisition and age, respectively).

ABIDE dataset

Functional connectivity data from the ABIDE [62] database were used to investigate the potential motion and center bias (as previously reported, e.g., by [13, 14] or [12]) when training models that aim to predict ASD diagnosis.

Imaging center and in-scanner motion (normalized mean framewise displacement) were statistically significantly associated with ASD diagnosis ($R^2 = 0.019$ and 0.028, respectively; $P < 0.001$ for both; see also Table 2). The model trained on the raw (unadjusted) connectivity features predicted diagnosis with a medium effect size ($R^2 = 0.126$, ROC AUC = 0.71, $P < 0.001$).

The partial confounder test revealed that the raw model was significantly biased for both age group and acquisition batch (both $P < 0.0001$; see first column in Fig. 7). Predictions for several sites (e.g., Carnegie Mellon University, University of Leuven, Social Brain Lab UMC Groningen) were severely miscalibrated, and higher motion was associated with a higher probability for ASD diagnosis.

Both feature regression and COMBAT seemed to significantly attenuate center bias, but with COMBAT, the partial confounder test still provided evidence for a significant residual bias (0.009, third columns of the first row in Fig. 7).

When trying to mitigate the effect of in-scanner motion (bottom row in Fig. 7), both confounder mitigation approaches seemed to effectively mitigate motion bias, as suggested by the partial confounder test ($P > 0.05$, middle and right panels in the bottom row of Fig. 7).

Both feature regression and COMBAT considerably improved the predictive performance when mitigating center effects (AUC = 0.71 without correction and 0.75 with both feature regression and combat). With both feature regression and COMBAT, however, the effort of mitigating motion effects happened at the cost of a drop in predictive performance (AUC = 0.69 and 0.70, for feature regression and COMBAT, respectively).

Discussion

The concept of conditional independence provides a straightforward framework for assessing confounding bias in predictive models, assuming that both the target variable and the potential confounder have been observed for the validation dataset. However, handling the nonnormal and/or nonlinear conditional dependencies often seen in predictive models [37, 38] (Supplementary Figs. S10–S11) poses a great challenge. In fact, as recently shown by Shah and Peters in their “no free lunch” theorem [35], it is effectively impossible to establish a *fully nonparametric* conditional independence test with a valid type I error control and a nontrivial power. Indeed, perhaps somewhat surprisingly, but not totally unexpectedly [31], partial correlation–like analogs of a widely used bivariate nonparametric test, like partial Spearman correlation, exhibit inflated type I errors even with slight violations of normality and/or linearity (as clearly demonstrated with simulated data in Fig. 3). While the magnitude of this problem may not be fully appreciated in case of predictive model diagnostics, such tests are, in general, poor choices for testing confounding bias in machine learning. Conditional independence-based confounding bias testing must, therefore, be designed so that its suitability for the particular problem may be judged easily.

These tests place no assumptions on the conditional distributions of the model output, ensuring valid model diagnostics even in cases of nonnormally and nonlinearly dependent predictions. This property distinguishes the approach from other alternatives as it guarantees a valid type I error control even in cases of nonnormally and nonlinearly dependent predictions (i.e., in cases where Pearson and Spearman partial correlations and many other methods fail).

The proposed tests are based on solid theoretical foundations, underpinned by mathematical proofs. The main purpose of the simulated and empirical experiments was, therefore, not to justify the validity of the approach but to (i) test the software implementation, (ii) estimate statistical power in various situations, and (iii) exemplify how the partial confounder test can be used with real experimental data. The validity of the type I error control and was confirmed by our simulations, even if both the predictions and the confounder are nonnormally and/or nonlinearly dependent on the target variable (except by extreme nonnormality). While different biomedical applications may consider different amounts of bias to be relevant, in most cases, it is possible to set an upper bound for confounding bias that is still tolerable in certain applications. The simulation results can serve as a basis for power calculations in these cases, in order to identify the necessary sample size for proper model diagnostics.

A characteristic example for the potential areas of applications is the novel field of population neuroscience, where applying predictive modeling and machine learning on large-scale functional neuroimaging data holds great potential for both revolutionizing our understanding of the physical basis of mind and delivering clinically useful tools for diagnostics or therapeutic decision-making [3, 5, 23, 25]. However, the presence of confounders that are typical for biomedical research (e.g., sample demographics, center effects) or specific to the data acquisition and processing approach (e.g., imaging artifacts) presents a great challenge to these efforts [29]. The usefulness of the proposed tests is demonstrated in 2 such examples, using the HCP [55] and the ABIDE [62] datasets.

In the case of the HCP dataset, the statistically significant age bias of the “raw” model for predicting fluid intelligence is in line with previous findings [18, 19] and could likely exaggerate to a serious bias when testing the model on data of participants outside of the—relatively narrow—age range of the HCP sample. In this case, the bias would likely significantly harm the out-of-sample generalizability of this model. The bias of the same model for acquisition batch can also be problematic, especially as it has not yet been thoroughly discussed in case of the HCP dataset. There can be manifold reasons for the observed acquisition bias. Fluid intelligence of the included participants might be, for instance, affected by a changing selection bias during participant recruitment (e.g., as a consequence of the HCP receiving an increasing degree of public interest during its course).

In the ABIDE dataset, neither the center bias nor the age bias is surprising in the case of the “raw” model, but both would be obviously severely problematic for a diagnostic biomarker candi-

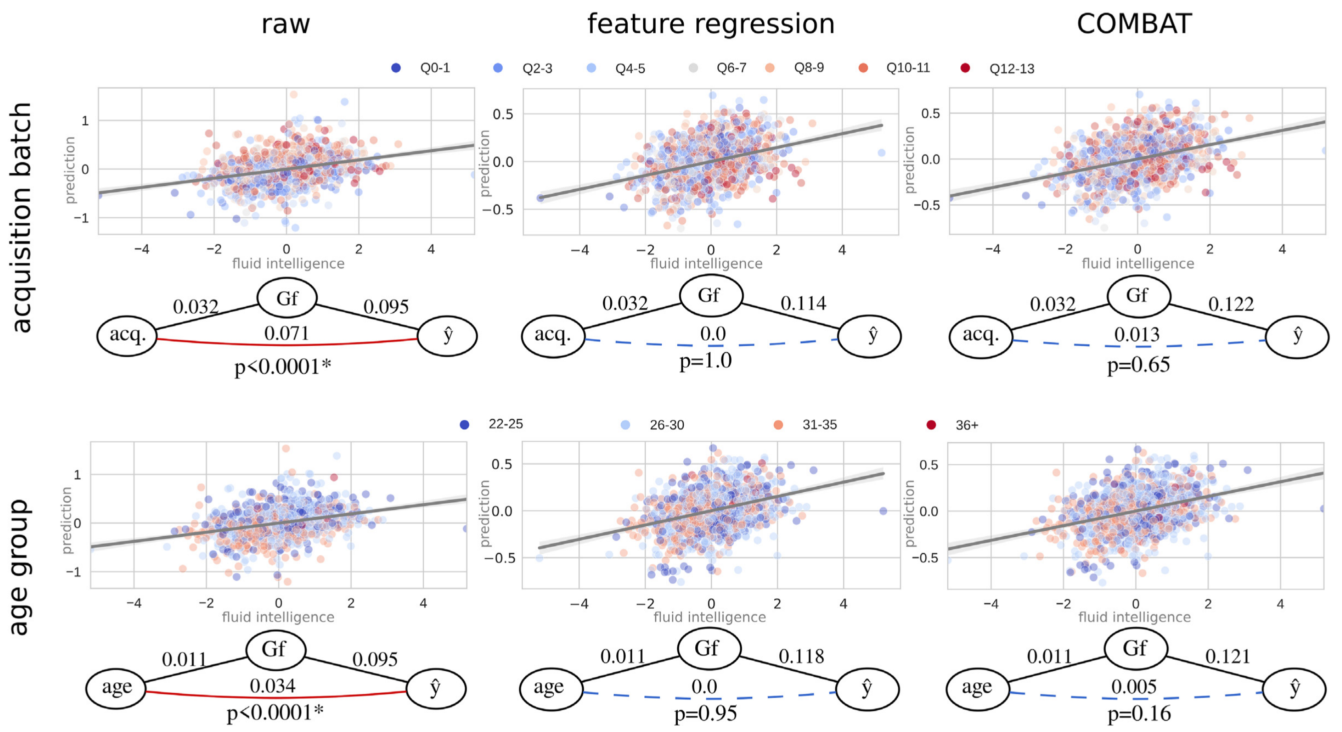


Figure 6: The partial confounder test reveals that acquisition batch and age bias in predictive models of fluid intelligence can be effectively attenuated by confounder mitigation approaches in the HCP dataset. Scatterplots and regression lines (with 95% confidence intervals) show the association of the observed (horizontals axis) and predicted (vertical axis) fluid intelligence scores with various confound regression strategies. Color-coding of the confounder variables (top: acquisition batch, bottom: age group, as shown by the corresponding legends) reveals confounding bias for both acquisition and age in the models trained on the raw data. This bias is robustly detected by the partial confounder test ($P < 0.0001$) and seems to be effectively mitigated by both feature regression and COMBAT. Relation between the observed (G_f) and predicted ($\hat{y}$) intelligence scores and the confounder variables is given on the graphs via R^2 values. Both confound mitigation techniques, but especially COMBAT, improve the predictive performance. Solid red line between the confounder and the prediction means significant confounding bias, whereas blue dashed line denotes that confounder testing provided no evidence for bias. P values are determined with the partial confounder test.

date of ASD. For instance, the model trained on the raw (unadjusted) features—depending on the calibration of the predicted class probabilities—might classify all participants from, for example, the Carnegie Mellon University center, as neurotypical control participants. Similarly, the models biased by motion—next to having questionable neuroscientific validity—might systematically fail in populations with a tendency for higher in-scanner motion (as known for many conditions, among others, attention-deficit/hyperactivity disorder [10] or Alzheimer's disease [9]).

The partial confounder test provided quantitative, statistically rigorous metrics for assessing the effectiveness of the investigated confounder mitigation techniques. In the HCP data, it revealed that both the acquisition bias and the age bias were very effectively removed by both feature regression and COMBAT ($P > 0.05$ for all). Given the high power of the test at the sample sizes of the HCP dataset ($N = 999$), any remaining confounding bias is most probably very safely negligible and well out of the range of practical relevance.

The confound mitigation approaches performed well in attenuating motion bias in the ABIDE dataset, as well, as no residual bias was detected by the proposed test. However, the success of COMBAT in eliminating motion bias is not to be taken without any objections. As COMBAT was originally developed for harmonizing effects of categorical variables (e.g., center or batch), its application for continuous confounder variables is not trivial. Inputting discretized versions of continuous variables into COMBAT might be suboptimal and raises further questions, for example, regarding the optimal number of bins used during the discretization.

Importantly, the partial confounder test revealed that the center bias of the classification in the massively multicenter ABIDE dataset was, although mitigated, not successfully removed by COMBAT. While determining the relevance of the remaining bias is out of the scope of this article, the example demonstrates the need for checking confounder bias even if state-of-the-art confounder mitigation approaches have been applied. If the proposed test provides evidence for residual confounding bias, the researcher might consider the use of another mitigation approach (e.g., feature regression in the given case) or the evaluation of confound-free performance (e.g., via "confound-isolating cross-validation") [29].

In sum, the application of the partial confounder test on the real data examples suggests that confounding bias must be always carefully investigated and reported in studies utilizing predictive modeling and machine learning as (i) variables as trivial as the date of the acquisition can cause significant confounding bias, and (ii) in certain situations, state-of-the-art confounder mitigation techniques may not provide sufficient mitigation of confounding bias, and (iii) unnecessary confounder correction may eliminate variance of interest. As the proposed test is a model-agnostic post hoc test, it can be used to benchmark different machine learning models and to further characterize already trained models in external validation samples, where a larger set of potential confounder variables is available (Supplementary Fig. S13). The partial confounder test can be considered a useful, objective benchmark to guide the search for a suitable confounder mitigation approach for every dataset.

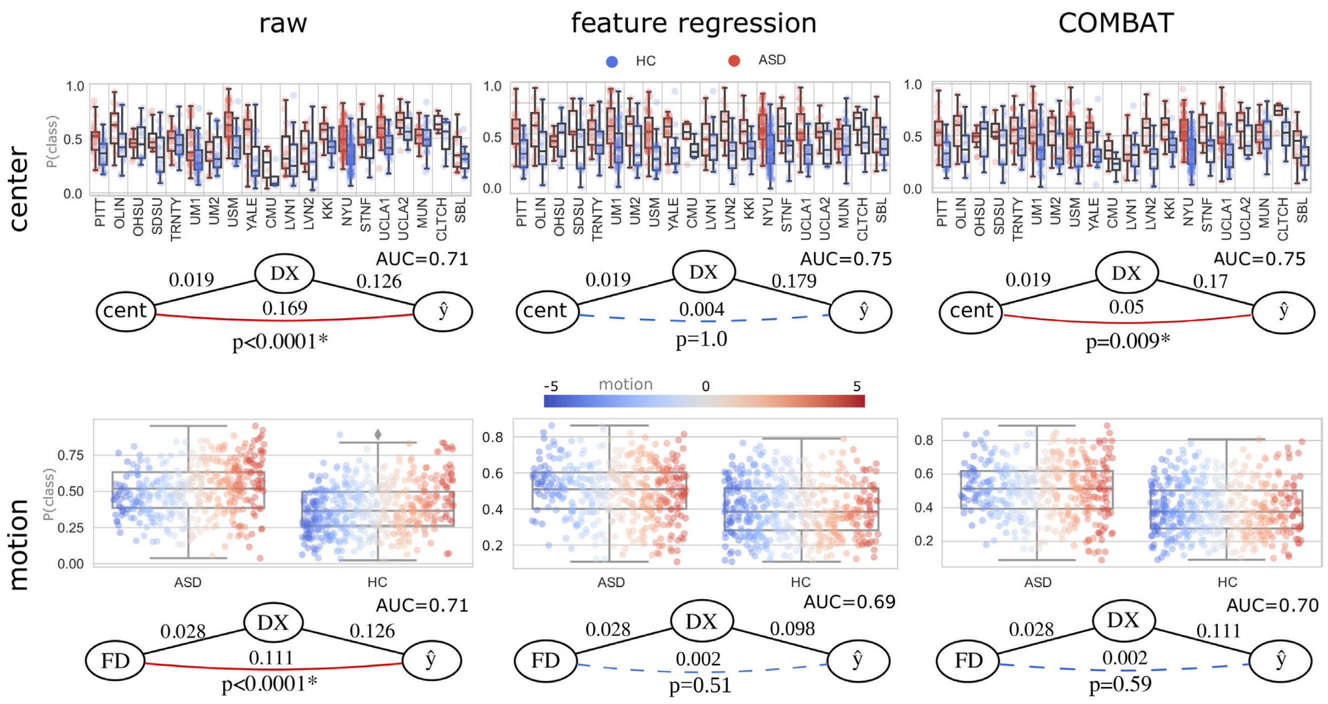


Figure 7: The partial confounder test identifies an efficient mitigation strategy for motion bias in predictive models of autism spectrum disorder and reveals residual center bias after COMBAT in the ABIDE dataset. Boxplots and points show the predicted class probabilities (0: Healthy Control (HC), 1: ASD), separately for the HC and ASD groups. In the top panel, predictions are plotted for each center separately. Color indicates the true diagnosis (DX). At the bottom plot, color indicates the normalized index of in-scanner motion (normalized FD). The proposed confounder test reveals significant center and motion bias in the model trained on the raw data ($P < 0.0001$). While both motion and center bias were effectively mitigated by both feature regression and COMBAT, the proposed partial confounder test revealed COMBAT was not able to fully remove center bias and resulted in significant "residual bias" ($P < 0.05$). Relation between the true ($\hat{y}$) and predicted diagnosis scores and the confounder variables is shown by the graphs as R^2 values. Solid red line between the confounder and the prediction means significant confounding bias, whereas blue dashed line denotes that confounder testing provided no evidence for bias. *P* values are determined with the partial confounder test.

Conclusion

The lack of rigorous statistical tests for confounding bias significantly hampers the development of predictive models in many fields of research, including population neuroscience, where handling confounding effects is especially challenging [23].

To fill this critical gap in predictive model development, here I proposed 2 novel tests, the *partial* and the *full confounder tests*, which probe the null hypotheses of "no confounding bias" and "full confounding bias," respectively. The tests are distinguished from alternative approaches by their robustness to nonnormally and nonlinearly dependent predictions, rendering them applicable with a wide variety of machine learning models. The tests have, moreover, a minimal computational overhead, as refitting the model is not required.

As demonstrated on functional brain connectivity-based predictive models of fluid intelligence and ASD, the tests can guide the optimization of confound mitigation strategies and allow quantitative statistical assessment of the robustness, generalizability, and neurobiological validity of predictive models in biomedical research. Given their simplicity, robustness, wide applicability, high statistical power, and computationally effective implementation (available in the Python package *mlconfound*; https://mlconfound.readthedocs.io), the partial and full confounder tests emerge as novel tools in the methodological arsenal of predictive modeling and may largely accelerate the development of clinically useful machine learning biomarkers.

Data Availability

Empirical analysis was based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium [55] (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657), funded by the 16 National Institutes of Health (NIH) institutes and centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University and the ABIDE consortium [62].

All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms. Python implementation of the "mlconfound" package is available on GitHub. All analysis code is available at GitHub and via the GigaScience database GigaDB [69].

Additional Files

Supplemental Figure S1. Heatmaps showing the positive rates of the "partial" confounder test, with categorical variables, normal conditional distribution, and linear dependence.

Supplemental Figure S2. Heatmaps showing the positive rates of the "full" confounder test, with numerical variables, normal conditional distribution, and linear dependence.

Supplemental Figure S3. Heatmaps showing the positive rates of the “full” confounder test, with categorical variables, normal conditional distribution, and linear dependence.

Supplemental Figure S4. Heatmaps showing the positive rates of the “partial” confounder test, with numerical variables, normal conditional distribution, and sigmoid dependence.

Supplemental Figure S5. Heatmaps showing the positive rates of the “partial” confounder test, with categorical variables, normal conditional distribution, and sigmoid dependence.

Supplemental Figure S6. Heatmaps showing the positive rates of the “full” confounder test, with numerical variables, normal conditional distribution, and sigmoid dependence.

Supplemental Figure S7. Heatmaps showing the positive rates of the “full” confounder test, with categorical variables, normal conditional distribution, and sigmoid dependence.

Supplemental Figure S8. Histogram of fluid intelligence score in the HPC dataset, before (left) and after (right) quantile transformation.

Supplemental Figure S9. Histogram of mean framewise displacement in the ABIDE dataset, before (left) and after (right) quantile transformation.

Supplemental Figure S10. Example of nonlinearity of model predictions as a consequence of regularization. The same 4 (simulated) features may result in nonlinear predictions as the regularization (alpha) of the Ridge model is increased. Model coefficients are shown above the lines connecting the features and the prediction. The full analysis is available at https://github.com/pni-lab/mlconfound- manuscript/blob/main/simulated/normalityandlinearityviolation.ipynb.

Supplemental Figure S11. Example of nonnormality of the conditional distributions $\hat{y}|y$ and $\hat{y}|c$. (A) Example from the analysis of the HCP dataset, as presented in the previous version of the manuscript. (B) No evidence of confounder bias with the partial confounder test. (C) Presumably false-positive observations by Pearson’s and Spearman’s partial correlations, due to invalid *P* values with nonnormal conditional distributions. Prediction target: age. Confounder: age. Confound mitigation: age regression. Nonnormality was frequently observed in the other cases, as well. The full analysis is available at https://github.com/pni-lab/mlconfound- manuscript/blob/main/empirical/supplement/check_assumptions.ipynb.

Supplemental Figure S12. In case of linearity and normality, the power of the proposed test is virtually equal to that of Pearson’s partial correlation. Blue: partial confounder test; orange: Pearson’s partial correlation. Boxplots are based on the simulation cases from Fig. 4 of the article.

Supplemental Figure S13. The partial confounder test can be used at any phase of model validation. The NYU site from the ABIDE dataset has been used as a “discovery sample” to train a model predicting ASD diagnosis. The partial confounder test found no evidence for motion bias. The finalized model has been externally validated in data from the University of Utah School of Medicine (USM). Next to the repeated testing of motion bias, the proposed test is used here for testing another potential confounder (fluid intelligence, G_f) and, additionally, to test if the model generalizes to the SRS (Social Response Scale). Source code available at https://github.com/pni- lab/mlconfound-manuscript/blob/main/empirical/supplement/external_validation.ipynb.

Abbreviations

ABIDE: Autism Brain Imaging Data Exchange; ASD: autism spectrum disorder; AUC: area under the curve; COMBAT: “combating batch effects” data harmonization approach; CPT: conditional permutation testing; DX: diagnosis; FD: framewise displacement; GAM: generalized additive model; G_f: fluid intelligence; HCP: Human Connectome Project; MCMC: Markov chain Monte Carlo; ROC: receiver operator curve.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Projektnummer 316803389–SFB 1280 and TRR 289 Treatment Expectation—Projektnummer 422744262.

Acknowledgments

I am thankful to Ulrike Bingel (University Hospital Essen, Germany) and Robert Englert (University Hospital Essen, Germany) for their valuable insights and comments on the manuscript. I show appreciation to the contributors of the Human Connectome Project and the Autism Brain Imaging Exchange study for collecting and sharing the quality data to researchers.

References

1. Vogt, N. Machine learning in neuroscience. *Nat Methods* 2018;**15**(1):33.
2. Kent, DM, Steyerberg, E, van Klaveren, D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ* 2018;**363**:k4245
3. Spisak, T, Kincses, B, Schlitt, F, *et al.* Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. *Nat Communications* 2020;**11**(1):1–12.
4. Walsh, I, Fishman, D, Garcia-Gasulla, D, *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* 2021;**18**:1122–27.
5. Woo, CW, Chang, LJ, Lindquist, MA, *et al.* Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci* 2017;**20**(3):365–77.
6. Obermeyer, Z, Powers, B, Vogeli, C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**(6464):447–53.
7. Mehrabi, N, Morstatter, F, Saxena, N, *et al.* A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 2021;**54**(6):1–35.
8. Prosperi, M, Guo, Y, Sperrin, M, *et al.* Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Machine Intell* 2020;**2**(7):369–75.
9. Rao, A, Monteiro, JM, Mourao-Miranda, J, *et al.* Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* 2017;**150**:23–49.
10. Eloyan, A, Muschelli, J, Nebel, MB, *et al.* Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Front Syst Neurosci* 2012;**6**:61.
11. Couvy-Duchesne, B, Ebejer, JL, Gillespie, NA, *et al.* Head motion and inattention/hyperactivity share common genetic influences: implications for fMRI studies of ADHD. *PLoS One* 2016;**11**(1):e0146271.
12. Gotts, SJ, Saad, ZS, Jo, HJ, *et al.* The perils of global signal regression for group comparisons: a case study of autism spectrum disorders. *Front Hum Neurosci* 2013;**7**:356.

13. Spisak, T, Jakab, A, Kis, SA, *et al.* Voxel-wise motion artifacts in population-level whole-brain connectivity analysis of resting-state FMRI. *PLoS One* 2014;**9**(9):e104947.
14. Spisak, T, Kincses, B, Bingel, U. Optimal choice of parameters in functional connectome-based predictive modelling might be biased by motion: comment on Dadi et al. *bioRxiv* 2019;710731. doi: https://doi.org/10.1101/710731.
15. Orban, C, Kong, R, Li, J, *et al.* Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS Biol* 2020;**18**(2):e3000602.
16. Cole, MW, Yarkoni, T, Repovš, G, *et al.* Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *J Neurosci* 2012;**32**(26):8988–99.
17. He, T, Kong, R, Holmes, AJ, *et al.* Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* 2020;**206**:116276.
18. Dubois, J, Galdi, P, Paul, LK, *et al.* A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Trans R Soc B Biol Sci* 2018;**373**(1756):20170284.
19. Lohmann, G, Lacosse, E, Ethofer, T, *et al.* Predicting intelligence from fMRI data of the human brain in a few minutes of scan time. *bioRxiv* 2021;435935. https://doi.org/10.1101/2021.03.18.435935.
20. Lwowski, B, Rios, A. The risk of racial bias while tracking influenza-related content on social media using machine learning. *J Am Med Inform Assoc* 2021;**28**(4):839–49.
21. Li, J, Bzdok, D, Holmes, A, *et al.* Not one model fits all: unfairness in RSFC-based prediction of behavioral data in African American. Helmholtz AI Kick-off Meeting. 2020;MunichGermany. https://orbi.uliege.be/handle/2268/245640.
22. Paulus, MP, Thompson, WK. Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology* 2021;**238**(5):1231–9.
23. Smith, SM, Nichols, TE. Statistical challenges in "big data" human neuroimaging. *Neuron* 2018;**97**(2):263–8.
24. Wachinger, C, Rieckmann, A, Pölsterl, S, *et al.* Detect and correct bias in multi-site neuroimaging datasets. *Med Image Anal* 2021;**67**:101879.
25. Nunes, A, Schnack, HG, Ching, CR, *et al.* Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Mol Psychiatry* 2020;**25**(9):2130–43.
26. Dukart, J, Schroeter, ML, Mueller, K, *et al.* Age correction in dementia—matching to a healthy brain. *PLoS One* 2011;**6**(7):e22193.
27. Abdulkadir, A, Ronneberger, O, Tabrizi, SJ, *et al.* Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. In: *International Workshop on Pattern Recognition in Neuroimaging*. 2014; IEEE, Tuebingen:Germany.
28. Johnson, WE, Li, C, Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**(1):118–27.
29. Chyzhyk, D, Varoquaux, G, Milham, M, *et al.* How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* 2022;**11**:giac014.
30. Dockès, J, Varoquaux, G, Poline, JB. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience* 2021;**10**(9):giab055.
31. Korn, EL. The ranges of limiting values of some partial correlations under conditional independence. *Am Stat* 1984;**38**(1): 61–2.
32. Bergsma, W. *Nonparametric testing of conditional independence by means of the partial copula. arXiv preprint* 2011;arXiv:1101.4607
33. Candès, E, Fan, Y, Janson, L, *et al. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection*. arXiv preprint arXiv:161002351 2016.
34. Peters, J, Bühlmann, P, Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc B* 2016;**78**(5):947–1012.
35. Shah, RD, Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Ann Stat* 2020;**48**(3):1514–38.
36. Berrett, TB, Wang, Y, Barber, RF, *et al.* The conditional permutation test for independence while controlling for confounders. *J R Stat Soc B* 2020;**82**(1):175–97.
37. García, S, Fernández, A, Luengo, J, *et al.* A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing* 2009;**13**(10):959.
38. Kristensen, SB, Sandberg, K. Is whole-brain functional connectivity a neuromarker of sustained attention? Comment on Rosenberg et al. (2016). bioRxiv 2017; 216697.
39. Neto, CE, Pratap, A, Perumal, TM, *et al.* A permutation approach to assess confounding in machine learning applications for digital health. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery 2019; New York: United States.
40. Ferrari, E, Retico, A, Bacciu, D. Measuring the effects of confounders in medical supervised classification problems: the Confounding Index (CI). *Artificial Intell Med* 2020;**103**:101804.
41. Southworth, LK, Kim, SK, Owen, AB. Properties of balanced permutations. *J Comput Biol* 2009;**16**(4):625–38.
42. Hemerik, J, Goeman, J. Exact testing with random permutations. *Test* 2018;**27**(4):811–25.
43. Dawid, AP. Conditional independence in statistical theory. *J R Stat Soc B* 1979;**41**(1):1–15.
44. Spirtes, P, Glymour, CN, Scheines, R, *et al. Causation, Prediction, and Search*. Cambridge, MA: MIT Press; 2000.
45. Fiedler, K, Schott, M, Meiser, T. What mediation analysis can (not) do. *J Exp Soc Psychol* 2011;**47**(6):1231–36.
46. Pitman, EJ. Significance tests which may be applied to samples from any populations. *Suppl J R Stat Soc* 1937;**4**(1):119–30.
47. Fisher, R. The Theory of Confounding in Factorial Experiments in Relation to the Theory of Groups *Annals of Eugenics*. 1942;**11**:341–53. Cambridge University Press.
48. Hastie, T, Tibshirani, R. Generalized additive models: some applications. *J Am Stat Assoc* 1987;**82**(398):371–86.
49. Bennett, B. Multiple regression analysis of binary and multinomial variates. *The Indian Journal of Statistics* 1966; **28**(3/4):301–4.
50. Jones, RH. Probability estimation using a multinominal logistic function. *J Stat Comput Simul* 1975;**3**(4):315–29.
51. Chambers, M, Dinsmore, TW. *Advanced Analytics Methodologies: Driving Business Value with Analytics*. Pearson Education; 2014.
52. Servén, D, Brummitt, C, Abedi, H. pyGAM: generalized additive models in Python(v0.8.0) Zenodo; 2018;https://doi.org/10.5281/zenodo.1476122
53. Campbell, M Karen, Donner, Allan Classification efficiency of multinomial logistic regression relative to ordinal logistic regression *Journal of the American Statistical Association* 1989;**84**(406):587–91.
54. Jones, MC, Pewsey, A. Sinh-arcsinh distributions. *Biometrika* 2009;**96**(4):761–80.
55. Van Essen, DC, Smith, SM, Barch, DM *et al.* The WU-Minn human connectome project: an overview. *Neuroimage* 2013;**80**: 62–79.

56. Glasser, MF, Sotiropoulos, SN, Wilson, JA, *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 2013;**80**:105–24.
57. Duncan, J, Seitz, RJ, Kolodny, J, *et al.* A neural basis for general intelligence. *Science* 2000;**289**(5478):457–60.
58. Beasley, TM, Erickson, S, Allison, DB. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet* 2009;**39**(5):580–95.
59. Pedregosa, F, Varoquaux, G, Gramfort, A, *et al.* Scikit-learn: Machine learning in Python. *J Machine Learn Res* 2011;**12**:2825–30.
60. Fortin, JP, Cullen, N, Sheline, YI, *et al.* Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 2018;**167**:104–20.
61. Hoerl, AE, Kennard, RW. Ridge regression: applications to nonorthogonal problems. *Technometrics* 1970;**12**(1):69–82.
62. Di Martino, A, Yan, CG, Li, Q, *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* 2014;**19**(6):659–67.
63. Dadi, K, Rahim, M, Abraham, A, *et al.* Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage* 2019;**192**:115–34.
64. Craddock, C, Benhajali, Y, Chu, C *et al.*, The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Neuroinformatics* 2013; Stockholm, Sweden. https://www.frontiersin.org/10.3389/conf.fninf.2013.09.00041/event_abstract.
65. Bellec, P, Rosa-Neto, P, Lyttelton, OC, *et al.* Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage* 2010;**51**(3):1126–39.
66. Huntenburg, J, Abraham, A, Loula, J, *et al.* Loading and plotting of cortical surface representations in Nilearn. *Res Ideas Outcomes* 2017;**3**:e12342.
67. Estève, L. Big data in practice: the example of nilearn for mining brain imaging data. *Scipy* 2015; Austin, Texas: United States.
68. Power, JD, Mitra, A, Laumann, TO, *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 2014;**84**:320–41.
69. Spisak, T. Supporting data for "Statistical quantification of confounding bias in machine learning models." *GigaScience Database* 2022. http://dx.doi.org/10.5524/102244.